Linked open data, expanded authority search capabilities, and synonym expansion

Final Proposal, April 23, 2012

Summary of changes since draft proposal (additions/changes in boldface):

- Updated paragraph 5, to add additional stakeholders/consultants to the project;
- Edited the list of tasks that the project will address, and then developed from those tasks a list of mid-project milestones that we can use to check our progress.

Proposal:

The Countway Library currently maintains on its website the Countway Digital Library, which consists of all the electronic books, journals, and databases, related to the interests of Harvard Medical faculty, researchers, and students. Every month over 30,000 of our researchers take advantage of the Digital Library to obtain access to our electronic materials, but as it is currently configured the Digital Library has limited search and discovery functions. We propose to improve the usefulness and discoverability of these materials by exposing bibliographic information as Linked Open Data, and then utilizing these new data points to enhance searches. We believe that this will make the data much more useful to our researchers, and it will also allow us to link our bibliographic data to other linked data sources, such as the Harvard Catalyst's Profiles, the Virtual International Authority File (VIAF), and eagle-i resources. We plan to work with the LibraryCloud team, to leverage their experiences with bibliographic information and Linked Open Data; they plan on incorporating our work into LibraryCloud and the Digital Public Library of America platform.

This project will also address a deficiency that currently exists in the library's public catalogs: the entry terms/see-references for subject headings and the variant forms of personal and corporate names are not included in searches, resulting in misleading search results. Take for example the following record in Aquabrowser [HOLLIS]:

Title:	Neoplasms of the liver
Author:	Kunio Okuda
Medical Subject Heading:	Liver Neoplasms

Performing a keyword search with the words "Kunio and Liver Neoplasms" will retrieve this record. However, if the keyword search is "Kunio and Liver Cancers", this record will **not** be retrieved as a result, despite the fact that "Liver Cancer" is an entry term within the MeSH authority record for the concept "Liver Neoplasm." Name searches in the public catalog also ignore see-references from name authority records. For example, a keyword search in Aquabrowser [HOLLIS] for "Samuel Clemens" results in 365 hits. A search with the words

“Mark Twain,” however, results in 2,099 hits, despite the fact that the name “Samuel Clemens” is a see-reference on the name authority record for Mark Twain.

By converting existing MARC records into a Resource Description Framework (RDF) representation, we will be able to link existing fields to external data sources that will enhance searches. For example, if the subject heading field were linked with MeSH, the entry term could be exposed, or if the names in the catalogue were linked with VIAF, pseudonyms and actual names could be interpreted as one person. To leverage this linkage, we will develop a search service that rewrites queries to improve discovery. Query rewriting is a process by which semantic understanding of a user’s query can be used to reconstruct the query in a way that improves information retrieval. Specifically, ontologies can be used to look up synonymy between terms and queries can be rewritten to increase accuracy and coverage.

The LibraryCloud team at the Harvard Law Library has experience with ingesting bibliographic metadata, exposing the data to the public, and researching some of the available ontologies that could be used for a schema. The Center for Biomedical Informatics (CBMI) team at the Countway Library has experience with storing and working with RDF, performing Extract Transform Load (ETL) on non-RDF data, query rewriting and providing Linked Open Data for consumption by external services. The Collections and Knowledge Management team at the Countway Library has experience with MARC bibliographic and authorities data, the MeSH thesaurus, and with the collection needs and search habits of Countway Library’s patrons. **Robin Wendler of OIS will act as a consultant, providing input throughout the course of the project and identifying other units on campus who are performing relevant work. Additionally, the Center for the History of Medicine and the Medical Heritage Project at Countway Library have signed on as observers. Their participation will allow us to expand the scope of the digital collections we work with, once a scalable production model has been achieved.**

We have already performed a small experiment with converting a small subset of MARC records into RDF using MarcEdit and loading it into an RDF repository that we instantiated for this purpose. While this experiment proved successful, we observed that a significant amount of the rich data in the MARC records was lost as well as a one-dimensional-ness in the resulting RDF due to the ontology that was used. The resulting RDF data and its ontology, however, was integrated nicely into one of CBMI’s existing software platforms. From this experiment and subsequent discussions, we plan on addressing the following tasks:

- Identify:
    - the ontologies that are best suited to our needs
    - an initial collection of electronic journals to be our test records
    - existing repository and software tools to store and manipulate RDF and expose it as Linked Open Data
    - other external data sources that could be linked to our test records

- Develop:
  - A light-weight ontological framework to incorporate additional ontology terms in the event no single ontology matches our needs
  - a convenient ETL method or tool to convert MARC data to RDF and establish the necessary linkage to external data sources.
  - a web-service that will access the RDF repository housing the converted RDF data and the new links to external data sources, and will perform query rewriting. **We intend to develop a query rewriting process that will also work with the public Hollis interface.**
  - an adapted or new search interface that incorporates the web-service

Mid-project milestones

Based on the tasks listed above, we have developed six mid-project milestones that can be used to measure our progress:

1. **Identify or develop the ontology or ontologies that we will use to frame the data.**
2. **Determine the RDF database and software tools to input, store, and expose the data.**
3. **Load the test records into the RDF database.**
4. **Identify the useful external data sources and link our test records to them.**
5. **Develop the query re-writing tool that will perform synonym and authority expansion.**
6. **Develop or identify a demonstrator user interface for searching.**

We believe that this project, although small in nature, will have large benefits to the Harvard community, and is therefore a good candidate for Library Lab. It represents an initial step at presenting library bibliographic information in a non-MARC and Semantic Web-friendly format. Since we are planning to develop a web-service, this enhanced functionality to searches can be expanded into other Harvard Library search interfaces. By successfully converting the MARC data to RDF and creating the links to external resources, we are providing a basis for future semantic capabilities that will enhance the users' experience with the library, such as synonym-based auto-suggest of certain fields in search interfaces. We believe that the project is scalable: since we will be developing an ETL tool to convert large amounts of MARC data, it will be relatively simple to streamline a process that converts the MARC data and creates links. Although the Center for Biomedical Informatics at Countway already uses Linked Open Data in some of its projects, to our knowledge this has not yet been attempted with library data. If our project succeeds, we will be able to share this model with other units on campus, and help move the library catalog past MARC.

Estimated Requested Funding:

No physical resources will be required for this project; software can run on existing CBMI/Countway servers. Countway Library will support project coordination activities. Project will request a partial chargeback for development efforts.

4 months x $89/hr = $54,400